УДК 57.088.1

СОЗДАНИЕ АЛГОРИТМА ОБРАБОТКИ ДАННЫХ, ВЫСОКОПРОИЗВОДИТЕЛЬНОГО СЕКВЕНИРОВАНИЯ ДНК *ESCHERICHIA COLI*

***Н.В. Жур**, магистрант 1 курса*
*Научный руководитель – **Д.Д. Жерносеков**, д.б.н., профессор*
Полесский государственный университет

Современные исследования в биологии и медицине все сложнее представить без использования математических методов. Постоянно увеличивающийся объем биологических данных, и развитие новых вычислительных технологий привели к формированию отдельного научного направления – биоинформатики. На сегодняшний день, методы биоинформатики особенно востребованы в современных генетических исследованиях, использующих новейшие технологии высокоплотной чип-гибридизации или секвенирования ДНК нового поколения

Развитие методов секвенирования нового поколения обеспечило высокую скорость, автоматизацию, качество и широкий спектр аналитических возможностей геномного анализа, в том числе для решения прикладных задач в биологии. Основные направления секвенирования нового поколения включают геномный анализ, направленное ресеквенирование геномов, секвенирование транскриптома, картирование ДНК-связывающих белков и анализ хроматина[2].

Инструменты информационных технологий позволяют обработать результаты научных исследований, произвести их статистический анализ, проверить достоверность гипотез, выявить биологические закономерности, придать результатам более наглядную, понятную всем форму.

На данный момент на рынке присутствуют несколько компаний, выпускающих устройства использующие методы секвенирования нового поколения. Одним из производителей этих устройств является компания Thermo Fisher Scientific, в последнее время их секвенаторы Ion Torrent стремительно набирают популярность благодаря своей дешевизне.

Существуют 2 основных подхода к сборке генома. Сборка генома de novo предполагает ассемблирование генома из коротких чтений без использования референсного генома, а также существует метод сборки геномов с использованием референсного генома [4,5]. Алгоритм сборки генома, как правило, включает в себя следующие этапы: 1. Секвенирование выделенных фрагментов ДНК или РНК. 2. Получение BAM файла и конвертация в Fastq. 3. Тримминг Fastq файла. 4. Сборка генома de novo. 6. Оценка качества контигов.

При исполнении каждого из этапов обработки данных была подобрана определённая утилита. Для контроля качества коротких чтений FastQC, фильтрации и тримминга trimmomatic. Одной из основных задач при сборке генома является выбор сборщика генома. При выборе программы были определены несколько основных параметров: объект исследования, платформа, используемая для секвенирования, принцип сборки используемый сборщиком, а также поддержка разработчиком и доступность. Так как для сборки использовались данные чтений генома *Escherichia coli* полученные при помощи Ion Torrent (данные чтений были взяты в The DNA Data Bank of Japan), то наилучшим образом под заданные параметры подошёл сборщик SPAdes(таблица 1)[1].

Таблица 1. – Параметры сборщика генома SPAdes

Объекты	Бактериальный геном
Платформы	Ion Torrent, PacBio, Oxford Nanopore, Illumina
Принцип сборки	Для сборки использует Графы де Брюйна
Поддержка	По сегодняшний день
Доступность	Бесплатно

При сборке были заданы три различные вариации показателей k-меров.

Эти показатели выбирались, опираясь на рекомендации разработчиков программы, а также научные публикации. После проведения трёх сборок был проведен анализ полученных сборок при помощи утилиты QUAST. Проведенная проверка показала, что показатели k-меров заданные при первой сборке (33,55,77) имели наилучшие показатели относительно сборок 2 и 3. Как видно из таблицы 2, первая сборка показала наименьшее количество контигов и их наибольший размер, также параметр N50 превышает таковое значение в двух остальных сборках более чем в 7 раз.

Таблица 2. – Результаты сборок, полученные при помощи QUAST

	Сборка 1	Сборка 2	Сборка 3
Количество контигов	1448	3360	3359
Наибольший контиг	27749	13846	15193
Содержание GC (%)	50,48	51,26	51,26
N50	5942	776	777

Анализ покрытия референсного генома показал, что сборка номер 1 смогла покрыть 80% генома в то время как 2 и 3 сборка – не более 40% (рис 1).

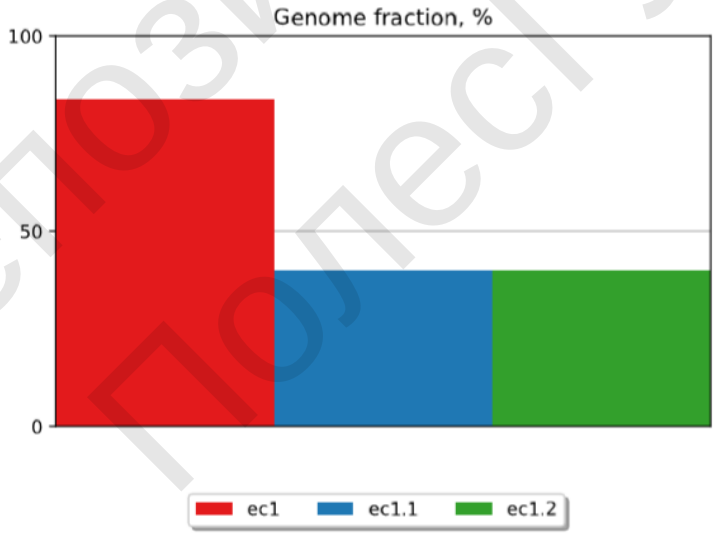

Рисунок – покрытие референсного генома *Escherichia coli*

Таким образом, нами были выделены основные стадии обработки данных секвенирования, осуществлён подбор программ для всех стадий сборки генома, а также определены оптимальные параметры для сборки генома *Escherichia coli.*

Список использованных источников

1. Bankevich, A. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing / A. Bankevich [et al.] // J Comput Biol. 2012, Vol. 19(5), P. 455–477.

2. Chiu, K.P. Next-generation sequencing and sequence data analysis. / K.P. Chiu // Bentham Science Publishers, 2015.

3. Denisov, G. Consensus generation and variant detection by Celera Assembler / G. Denisov [et al.] // BIOINFORMATICS. 2008, Vol. 24(8), P. 1035–1040.

4. Silva, G.G. Combining de novo and reference-guided assembly with scaffold_builder/ G.G. Silva [et al.] // Source Code Biol Med. 2013, Vol. 8(1), P. 23.

5. Zerbino, D.R. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler / D.R. Zerbino [et al.] // PLoS One. 2009, Vol. 4(12), P. 840.

6. Zhong, C. GRASP: guided reference-based assembly of short peptides / C. Zhong [et al.] // Nucleic Acids Res. 2015, Vol. 43(3), P. 18.